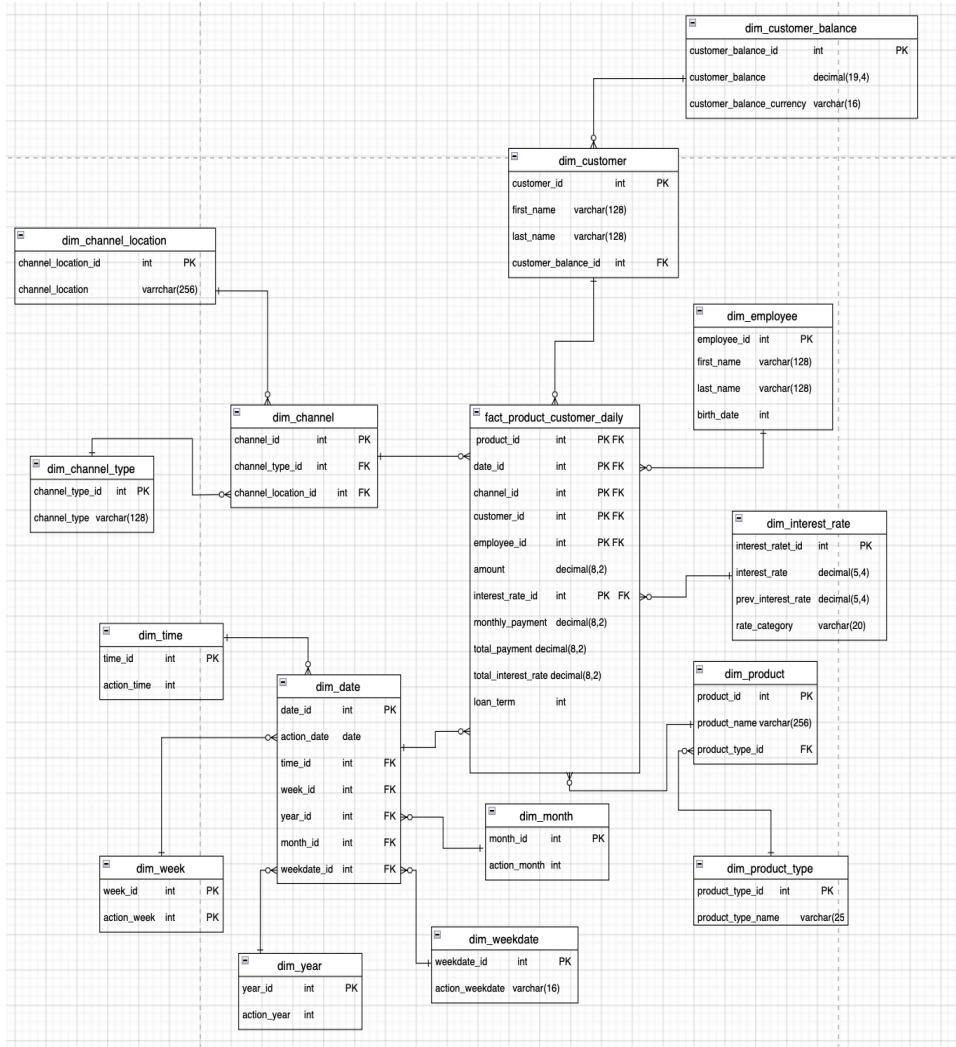


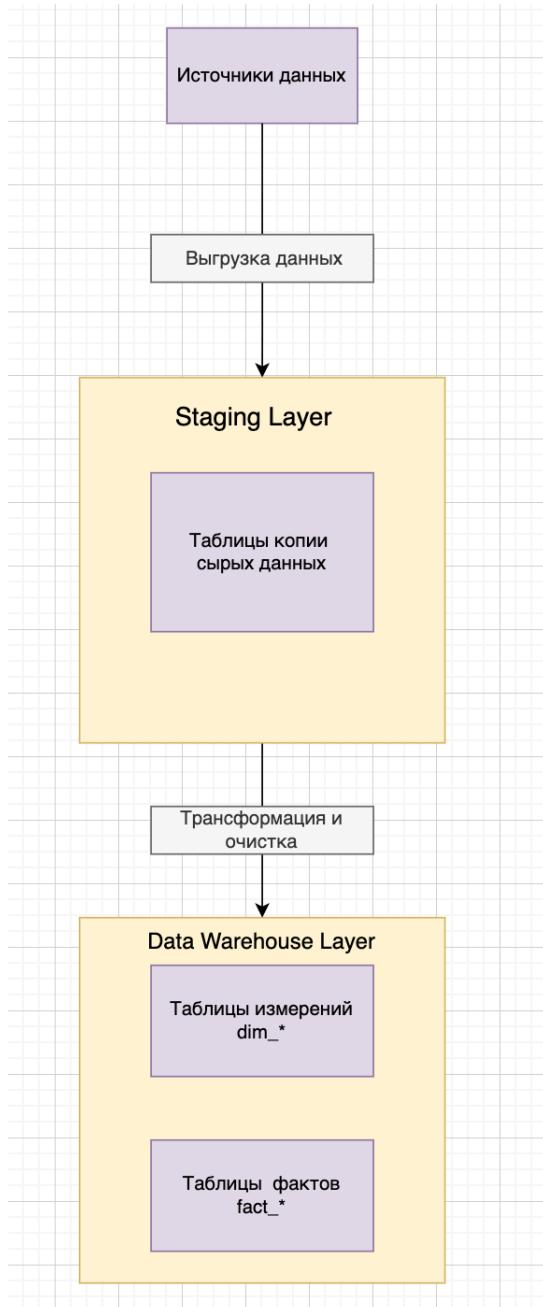
## Этап 3. Проектирование потоков данных и ETL-архитектуры

Выполнил:  
Воробьев А.И.

ER диаграмма ХД



## N1. Общая архитектура потока данных:



## Детальное описание этапов потока данных:

### 1. Источники данных (Data Sources)

данные поступают из различных операционных систем (OLTP):

- CRM-система: Откуда берутся данные о клиентах (customer), сотрудниках (employee) и каналах продаж (channel).

- Финансовая система / ядро банка: Откуда берутся данные о продуктах (product), балансах (customer\_balance) и финансовых операциях (факты о продажах, платежах).
- Логи (Logs): Откуда берутся данные о времени и датах действий (хотя измерений времени обычно создаются искусственно в ETL).

## 2. Промежуточная область (Staging Area)

Это область, куда данные попадают сразу из источников **в сыром виде**, без преобразований.

- **Цель:** Изолировать источники от нагрузки DWH, обеспечить быстрое извлечение и служить точкой восстановления в случае сбоя трансформации.
- **Структура:** Таблицы в Staging повторяют структуру таблиц-источников 1:1. Данные могут добавляться инкрементально или полностью перезаписываться каждый раз.

## 3. Трансформация и загрузка в DWH (Transformation & Load)

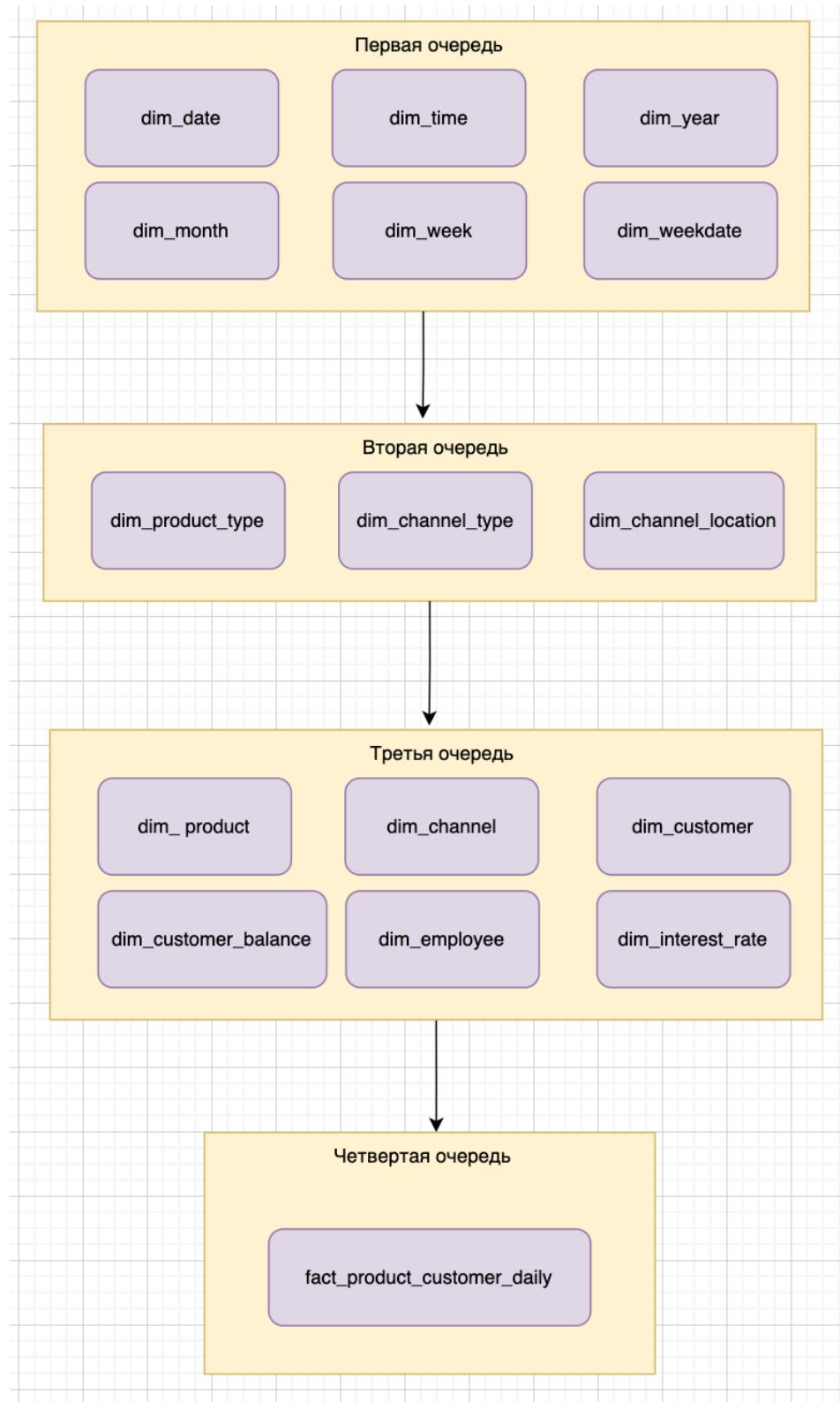
### a) Загрузка таблиц измерений (Dimensions):

- **Тип "Снежинка" (SCD) Type 1 (Перезапись истории):** Для атрибутов, где история изменений не важна. Например, dim\_channel\_location – если изменилось название местоположения, старое можно перезаписать.
- **Тип "Снежинка" (SCD) Type 2 (Сохранить историю):** Для атрибутов, где история изменений критична. Например, dim\_customer\_balance – необходимо знать, как баланс клиента менялся во времени. Добавляются поля start\_date, end\_date, is\_current.

### b) Загрузка таблицы фактов (Facts):

- **Процесс:** Каждая строка из сырых данных о продажах (напр., stg\_sales) проходит через серию подзапросов или JOIN'ов для замены "натуральных ключей" (например, имени продукта, даты) на **суррогатные ключи (ID)** из соответствующих таблиц измерений.
- **Даты:** Для замены даты на date\_id используется связка таблиц dim\_date, чтобы найти корректный ID по полной дате.

## N2. Порядок загрузки данных



1. **Календарные измерения** (dim\_date, dim\_time, dim\_year, dim\_month, dim\_week, dim\_weekdate) • Эти таблицы являются базовыми справочниками, которые не зависят от других источников. • Факт и остальные измерения используют их surrogate key (date\_id) для привязки записей. Поэтому они загружаются первыми, чтобы при загрузке фактов уже существовали все ключи дат.

2. **Справочные измерения без истории** (SCD1): dim\_product\_type, dim\_channel\_type, dim\_channel\_location • Эти таблицы содержат устойчивые классификаторы (типы продуктов, типы каналов, локации). • Они почти не меняются и не зависят от других сущностей. • Основные измерения (dim\_product, dim\_channel) используют их как FK. Их загружаем следующими - чтобы основные измерения имели готовые ссылки.

3. **Основные измерения (с историей – SCD2)**: dim\_product, dim\_channel, dim\_customer, dim\_customer\_balance, dim\_employee, dim\_interest\_rate • Эти таблицы описывают основные бизнес-сущности (клиенты, продукты, каналы, сотрудники, ставки). • Они зависят от справочников (например, dim\_product → dim\_product\_type, dim\_channel → dim\_channel\_type, dim\_channel\_location). • Для них используется стратегия SCD2, поэтому они должны быть загружены до фактов, чтобы факт можно было связать с актуальными surrogate keys. Поэтому на этом шаге формируется полный набор измерений, на который можно опираться при загрузке фактов.

4. **Факт** (fact\_product\_customer\_daily) • Факт - это «центральная таблица», которая хранит количественные показатели (сумма кредита, платежи, срок, ставка). • Он ссылается на все ключевые измерения (клиент, продукт, канал, сотрудник, ставка, дата). • Если измерения ещё не загружены, то факт нельзя корректно загрузить (FK будут отсутствовать). Факт загружается последним - после того как подготовлены все измерения.

N3

## Поток данных 1: Измерение dim\_date

- **Источник:** Не операционная система. Эта таблица заполняется искусственно.
- **Способ извлечения:** Генерация с помощью скрипта (SQL, Python) или специализированного ETL-инструмента (например, инструмент "Date Dimension" в SSIS или dbt).
- **Преобразования (Transform):**
  - Генерация диапазона дат (например, с 2000-01-01 по 2050-12-31).

- Вычисление атрибутов даты на основе каждой сгенерированной даты:
  - `action_date` (сама дата)
  - `day_of_week` (1-7), `day_of_month`, `day_of_year`
  - `is_weekend` (True/False)
  - `week_id` (номер недели по году)
  - `month_id` (номер месяца)
  - `year_id` (год)
  - `weekdate_id` (ссылка на `dim_weekdate`, например, 1=Monday)

Целевое поле (Target)	Источник (Source) / Преобразование	Описание
<code>date_id</code> (PK)	<code>GENERATE_SURROGATE_KEY()</code>	Суррогатный ключ, генерируется автоматически.
<code>action_date</code>	<code>date</code>	Сгенерированная дата.
<code>week_id</code>	<code>week_id</code>	Рассчитанный номер недели
<code>year_id</code>	<code>year_id</code>	Рассчитанный год
<code>month_id</code>	<code>month_id</code>	Рассчитанный номер месяца
<code>weekdate_id</code>	<code>weekdate_id</code>	Рассчитанный номер дня недели

## Поток данных 2: Измерение dim\_product

- **Источник:** Таблица products в OLTP-системе (финансовое ядро банка).
- **Способ извлечения (Extract):**
  - **Первая загрузка:** Полный выгрузка (FULL DUMP).
  - **Последующие загрузки: Инкрементальный по дате изменения** (INCREMENTAL LOAD). В таблице-источнике должен быть столбец last\_modified\_date. Выгружаются только строки, где last\_modified\_date > даты последней успешной загрузки.
- **Преобразования (Transform):**
  - **Очистка:** Приведение product\_name к единому регистру (верхнему). Удаление лишних пробелов.
  - **Обогащение:** Добавление поля product\_type\_id через JOIN с таблицей product\_types из источника или с уже загруженной dim\_product\_type.
  - **Дедубликация:** Проверка на уникальность по натуральному ключу (например, product\_code из источника).
  - **Обработка медленно меняющихся измерений (SCD):** История изменений важна **Type 2 (Создание новой версии)**

Целевое поле (Target)	Источник (Source) / Преобразование	Описание
product_id (PK)	GENERATE_SURROGATE_KEY()	Суррогатный ключ
product_name	TRIM(UPPER(source.product_name))	Очищенное название продукта
product_type_id (FK)	SELECT pt.product_type_id FROM dim_product_type pt WHERE pt.product_type_name = source.product_type_name	JOIN к целевой размерности для получения суррогатного ключа типа продукта

## Поток данных 3: Факты fact\_product\_customer\_daily

- **Источник:** Таблица sales или contracts в OLTP-системе.
- **Способ извлечения (Extract):**

- **Инкрементальная загрузка по дате операции.** Выгружаются все продажи за день, прошедший после последней загрузки (WHERE `sale_date = 'YYYY-MM-DD'`). Это ежедневная (daily) процедура.
- **Идеальный вариант: CDC (Capture Data Change)** на основе логов транзакций, чтобы захватывать все изменения в режиме, близком к реальному времени.
- **Преобразования (Transform):**
  - **Соединение с измерениями (Lookup):** Это **самое важное преобразование** для фактов. Для каждой строки продажи необходимо найти соответствующие суррогатные ключи во всех связанных измерениях.
    - Найти `product_id` по коду продукта из источника.
    - Найти `customer_id` по ID клиента из источника.
    - Найти `employee_id` по ID сотрудника из источника.
    - Найти `channel_id` по названию канала из источника.
    - **Найти `date_id` по дате продажи (`sale_date`)**. Это ключевой шаг.
  - **Валидация:** Отсеивание записей, для которых не найдены ключи в измерениях (например, поступила продажа по неизвестному продукту). Такие записи отправляются в таблицу ошибок для последующего разбора.
  - **Агрегация (опционально):** Если в источнике данные приходят на уровне транзакции, а в целевую таблицу нужно положить ежедневные итоги (что маловероятно, судя по названию таблицы), то проводится агрегация по ключам измерений.

Целевое поле (Target)	Источник (Source) / Преобразование	Описание
<code>product_id</code> (FK)	<code>Lookup('dim_product', source.product_code)</code>	Поиск суррогатного ключа продукта.
<code>product_id</code> (FK)	<code>Lookup('dim_date', source.sale_date)</code>	Поиск суррогатного ключа даты.
<code>channel_id</code> (FK)	<code>Lookup('dim_channel', source.channel_name)</code>	Поиск суррогатного ключа канала.
<code>customer_id</code> (FK)	<code>Lookup('dim_customer_balance')</code>	Поиск суррогатного ключа

	', source.customer_id)	клиента.
employee_id (FK)	Lookup('dim_employee', source.employee_id)	Поиск суррогатного ключа сотрудника.
amount	source.amount	Прямое отображение
interest_rate_id	source.interest_rate_id	Прямое отображение.
monthly_payment	source.monthly_payment	Прямое отображение.
total_payment	source.total_payment	Прямое отображение.
total_interest_rate	source.total_interest	Прямое отображение.
loan_term	source.loan_term	Прямое отображение.